# Similarité surfacique et similarité sémantique dans des cas cliniques générés

Nicolas Hiebel, Olivier Ferret, Karën Fort, Aurélie Névéol

# Similarité surfacique et similarité sémantique dans des cas cliniques générés

Nicolas Hiebel[1] Olivier Ferret[2] Karën Fort[3] Aurélie Névéol[1]
(1) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
(2) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
(3) Sorbonne Université / Université de Lorraine, CNRS, Inria, LORIA, 54506, Vandœuvre-lès-Nancy, France
[1]prenom.nom@lisn.upsaclay.fr, [2]olivier.ferret@cea.fr,
[3]karen.fort@loria.fr

MOTS-CLÉS : Génération, Similarité, Texte clinique, Texte synthétique, Français.

KEYWORDS: Generation, Similarity, Clinical Text, Synthetic Text, French.

**Contexte** La disponibilité restreinte des documents cliniques est un frein à la recherche en traitement automatique de la langue dans le domaine médical. Les corpus cliniques dont l'accès est relativement facile en français (E3C (Magnini *et al.*, 2020), CAS (Grabar *et al.*, 2018)) ne sont pas tout à fait représentatifs des documents confidentiels présents dans les hôpitaux. Le partage des connaissances au sein de la communauté scientifique est compliqué. Aucune reproductibilité n'est possible, tout comme les comparaisons avec d'autres méthodes / données. Une piste de création de ressource partageable en substitut des données confidentielles est la génération de données similaires à ces données privées. Cela pourrait permettre à des personnes ayant accès à un corpus privé de générer un corpus librement distribué à partir du premier. En partageant la méthode de génération, il serait également possible de reproduire l'expérience sur d'autres données confidentielles. La mise à disposition des données générées donnerait alors à la communauté scientifique un terrain de test, de comparaison, de discussion et d'entraide dans la recherche en TAL biomédical.

Les métriques existantes d'évaluation automatique de la génération de texte reposent majoritairement sur des mesures de similarité avec des références : une réponse idéale dans un contexte donné. Cette référence est comparée avec la ou les hypothèses du système (BLEU (Papineni *et al.*, 2002), ROUGE (Lin, 2004), BERTScore (Zhang *et al.*, 2020)). Nous présentons ici des alternatives à ces mesures pour évaluer une génération ouverte dans le domaine médical.

**Objectifs** L'évaluation de la génération est multidimensionnelle. Il est nécessaire d'observer la qualité de la langue dans les données générées, ainsi que son adéquation avec la langue des données sources. Dans cette étude, nous nous concentrons sur l'évaluation de la similarité entre des données médicales réelles et des données synthétiques générées à partir de ces données réelles selon deux critères complémentaires : la similarité surfacique et la similarité sémantique. Ces similarités sont importantes à estimer : une trop grande proximité pourrait signifier une copie d'information dans le corpus synthétique, représentant un potentiel risque de fuite d'information confidentielle. Inversement, il ne faut pas que les données synthétiques s'éloignent trop des données réelles pour que leur étude soit pertinente. Il y aura donc forcément des ressemblances. Nous cherchons par conséquent à repérer les éléments identiques ou ressemblants dans les deux jeux de données à l'aide de mesures de similarité,

en essayant de différencier les similarités traduisant un risque et les similarités acceptables.

**Méthodologie** Pour cette expérience, nous utilisons le corpus E3C, un corpus multilingue de documents biomédicaux librement disponible dans lequel nous sélectionnons les cas cliniques en français. Nous le désignerons par $\texttt{E3C}_{FR}$. Nous générons quatre corpus de données synthétiques de taille similaire au corpus $\texttt{E3C}_{FR}$ à l'aide de quatre configurations de modèles de langue génératifs pré-entraînés et adaptés (*fine-tune*) aux données cliniques de $\texttt{E3C}_{FR}$. Les modèles sont entraînés à générer des documents entiers en encadrant le texte généré par des balises de début et de fin du document. Nous étudions la similarité entre les données synthétiques générées et les données réelles à deux niveaux. Au niveau lexical, en observant les recouvrement de ngrammes entre les deux corpus, ce qui nous permet d'estimer la quantité de ngrammes que le modèle génératif a observé à l'entraînement et généré à l'identique à l'inférence. En complément, une comparaison avec un autre corpus de cas cliniques en français est ajoutée avec le corpus CAS. Deuxièmement, nous étudions la proximité sémantique entre les phrases des corpus. Pour cela, nous calculons les plongements des phrases des corpus à l'aide de l'outil SENTENCE-BERT (Reimers & Gurevych, 2019). Le modèle SENTENCE-BERT est ajusté sur le corpus de paires de phrases cliniques annotées en similarité CLISTER (Hiebel *et al.*, 2022) pour s'adapter au domaine clinique. Nous utilisons ensuite les plongements de phrases des deux corpus pour calculer une matrice de similarité entre les phrases des corpus, ce qui permet de récupérer pour chaque phrase du corpus généré les phrases les plus similaires dans le corpus réel avec les scores de similarité associés. C'est à partir de ces scores que nous essayons de déterminer les phrases et documents potentiellement problématiques du point de vue de la confidentialité.

**Résultats** Le tableau 1 présente les taux de recouvrement de ngrammes entre des corpus générés dans des configurations différentes et $\texttt{E3C}_{FR}$. Nous avons ajouté la comparaison avec le corpus CAS. On remarque que le corpus CAS présente plus d'unigrammes en commun avec le corpus $\texttt{E3C}_{FR}$. En revanche, il présente moins de séquences longues (8grammes) en commun avec ce même $\texttt{E3C}_{FR}$ que plusieurs corpus générés. On remarque également des différentes importantes entre les différents corpus générés. Le tableau 1 donne par ailleurs la similarité sémantique moyenne des phrases des corpus générés (et de CAS), calculée comme décrite à la section précédente. On observe ainsi que le corpus $\texttt{Bloom}_{E3C+T}$ présente une similarité moyenne plus élevée que les autres et le corpus $\texttt{LLF}_{E3C}$ la moins élevée, proche même de la proximité obtenue sur le corpus réel CAS.

| Corpus | 1gramme | 4gramme | 8gramme | Sim Sém |
|---|---|---|---|---|
| $\texttt{Bloom}_{E3C}$ | 0,16419 | 0,00368 | 0,00011 | 0,561 |
| $\texttt{Bloom}_{E3C+T}$ | 0,13887 | 0,00531 | 0,00020 | 0,585 |
| $\texttt{LLF}_{E3C}$ | 0,11740 | 0,00447 | 0,00023 | 0,531 |
| $\texttt{LLF}_{E3C+T}$ | 0,11935 | 0,00505 | 0,00013 | 0,557 |
| CAS | 0,20373 | 0,00899 | 0,00013 | 0,530 |

TABLEAU 1 – Comparaison entre les différents corpus générés et $\texttt{E3C}_{FR}$. Nous avons ajouté la comparaison entre $\texttt{E3C}_{FR}$ et CAS. Sim Sém = Similarité Sémantique moyenne.

**Conclusion** Ce travail donne un premier aperçu de la similarité surfacique et sémantique de corpus cliniques générés et naturels, montrant notamment que ces deux plans ne sont pas strictement corrélés.

# Références


GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, p. 122–128, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5614.

HIEBEL N., FERRET O., FORT K. & NÉVÉOL A. (2022). CLISTER : A corpus for semantic textual similarity in French clinical narratives. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 4306–4315, Marseille, France : European Language Resources Association.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

MAGNINI B., ALTUNA B., LAVELLI A., SPERANZA M. & ZANOLI R. (2020). The E3C Project : Collection and Annotation of a Multilingual Corpus of Clinical Cases. In J. MONTI, F. DELL'ORLETTA & F. TAMBURINI, Éds., *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 de *CEUR Workshop Proceedings* : CEUR-WS.org.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, p. 3980–3990 : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1410.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*, online.